Mechanically separated HAM and SPAM

May 27, 2015

Overview and Motivation

- Project: Use machine learning to predict whether emails are SPAM or not (“HAM”)

- Data: a subset of the Enron email corpus and a curated set of SPAM email (with malicious components removed)

    - http://www.aueb.gr/users/ion/data/enron-spam/

- Motivation: want to learn how to “quantify” text in order to apply machine learning techniques to extract information from large amounts of documents

Sample size

Data	Ham	Spam	Total
Raw data	19,088	32, 988	52,076
“Clean” data	18,962	22,006	40,968
Preprocessed data (w/ nulls dropped)	18,657	21,270	39,927

Adding structure to data

- Use python's **email** module to separate the email's body (text) from the email's metadata (headers, such as 'To', 'From', 'Subject', and 'Encoding')

- Use **beautifulsoup** to strip out all html tags that might be included in the email

Features

Currently, my model is built around only a few features of the possible features that can be extracted from this data:

- the document term matrix associated with the email text (p = 181,870)
- the document term matrix associated with the email subject (p = 19,435)

Initial results

(features = email text with replies/forwards included)

Model	Accuracy score	ROC-AUC score
Naive Bayes (w/o stop words removed)	0.987177	0.997176
Naive Bayes (w/ stop words removed)	0.987177	0.987177
Logistic Regression (w/o stop words removed)	0.985774	0.997331
Logistic Regression (w/ stop words removed)	0.984472	0.997101

Initial results

(features = email subject line)

Model	Accuracy score	ROC-AUC score
Naive Bayes (w/o stop words removed)	0.932779	0.980713
Naive Bayes (w/ stop words removed)	0.931877	0.981604

Confusion matrix

Model	TN	FP	FN	TP
Naive Bayes (w/o removing stop words)	4,509	63	65	5,345
Naive Bayes (w/ removing stop words)	4,505	67	61	5,349
Logistic Regression (w/o removing stop words)	4,563	110	32	5,277
Logistic Regression (w/ removing stop words)	4,549	124	31	5,278

Ex: false positives

- Dear Ms Kitchen You are a strong contender to be on Fortune s list of the Most Powerful Women in Business this year Please take a moment to make our Powerful Women issue as engaging and fun for our magazine readers and website users as it has been for us By answering the following questions you can help us to get a better sense of you and your busy life...

- Energy Info Source is privileged to make available to you a free sample issue of its Bi-weekly Transmission Update Report attached This report contains the latest ISO RTO Utility and Merchant Transmission news ...

- For my fellow travelers who will be going to Spain Italy Seattle and Pennsylvania Interesting travel web site ...

Ex: false negatives

- Doctors Use This Too Stay hard for straight incrrease si.ize and staa_mina with one piilll En+terr here http vsale I would come home late in the night and would get out early in the morning ...

- Final Notice Hi I sent you an email last week and need to confirm everything now Please read info below and let me know if you have any questions We are accepting your mortgage application …

- Good day I tried to call your three time but your phone is not available I think you did a mistake during filling the form Anyway your mortgagge request was appproved with please reenter your info here and we will start ASAP http Thank you Luke MCPKVL

Areas for further work

Creation of new features, such as:

- count number of links in email
- ratio of uppercase to lowercase characters
- number of non-alphanumeric characters
- length of email, subject, etc.
- timestamp (hour, day of the week, etc.)
- including n_grams

Creation of ensemble method to combine of naive bayes on email body and subject, and logistic regression (or others) on derived features listed above.

Tuning the model

Naive Bayes (w/o stop words removed)	Predict HAM	Predict SPAM
Actual HAM	4,509	63
Actual SPAM	65	5,345

Conclusions

- Naive Bayes is a very powerful tool
- Even initial runs of the models were returning accuracy rates greater than 98%
- I was only limited by my ability with work with text, rather than by the amount of information contained in the emails

Advice and comments always appreciated

https://github.com/jw-ml/dat5_spam-filter